International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 25s,2025
https://theaspd.com/index.php

# Distinguishing Ai Generated AND Human Crafted Phishing Emails: A Multi-Modal Machine Learning Approach WITH Adversarial Robustness Assessment

**Arnemie B. Gayyed[1], Natividad B. Concepcion[2]**
[1] College of Information Technology and Computer Science, University of the Cordilleras, Baguio City, Philippines.
[2] College of Information Technology and Computer Science, University of the Cordilleras, Baguio City, Philippines.
**Email:** [1]abgayyed@uc-bcf.edu.ph, [2]nbconcepcion@uc-bcf.edu.ph
**Orchid Id number:** [1]0009-0005-1523-9238, [2]0000-0001-6289-4909

***ABSTRACT:***
*AI-generated phishing emails pose a significant and evolving cybersecurity threat, rendering conventional detection techniques increasingly insufficient. This challenge reveals a critical research gap, the absence of a comprehensive model capable of accurately identifying sophisticated, AI-powered deceptive tactics. This study proposes a novel multi-modal machine learning framework to classify phishing emails, distinguishing between AI and human origins. A careful integration of diverse features, including textual patterns, content elements like HTML and embedded links, and vital metadata such as sender authentication. The researchers assembled a balanced dataset comprising 2780 human-generated emails from the Nazario corpus and an equivalent number of GPT-4 AI-generated emails. Results demonstrate the high effectiveness of all three feature modalities. Multi-modal models achieved impressive classification performance, with accuracy soaring to 100%. Metadata alone proved exceptionally powerful, yielding near-perfect detection with just 10 features. AI-generated emails showed distinct differences like more punctuation, fewer images, simpler HTML, while human-crafted ones featured longer URLs and more interactive elements. Despite excellent typical performance, models like Logistic Regression proved highly vulnerable to adversarial attacks, with accuracy dropping from 100% to 6% at $\epsilon$=0.5. In summary, this research provides empirical insights by establishing a robust multi-modal framework and critically examining its resilience against adversarial manipulations. These findings underscore the urgent need for smarter, multi-layered cybersecurity defenses to proactively counter escalating AI-driven threats.*

***KEYWORDS***: *AI-generated phishing email, cybersecurity, multi-modal features, adversarial robustness*

## 1) INTRODUCTION:

Despite growing efforts in phishing detection, existing models remained inadequate in identifying the origin of increasingly sophisticated attacks—particularly those generated by Artificial Intelligence (AI). Most systems focused on detecting malicious content or intent but failed to determine whether a phishing email was authored by a human or generated by an AI model. This origin classification gap was critical, as AI-generated phishing messages exhibited distinct linguistic, structural, and behavioral traits that evaded traditional filters. Without robust origin-aware frameworks, organizations remained vulnerable to novel attack strategies that exploited generative models' fluency, adaptability, and scale [1][2][3].

The urgency of this problem was underscored by global trends. In 2023, phishing was the most reported cybercrime, with approximately nine million cases and 86% of companies experiencing phishing attempts. Financial institutions were disproportionately targeted, accounting for 27.7% of attacks according to statista. AI-generated phishing emails—often crafted using models like GPT-4—became increasingly realistic, with click-through rates ranging from 30% to 44%, and even higher when combined with deception-enhancing techniques such as the V-Triad [3][4][5]. These attacks were not only more convincing but also more scalable, automating the entire phishing lifecycle from target selection to message distribution [6]. Despite growing awareness, the volume and sophistication of AI-enabled phishing often exceeded human detection capabilities [7].Traditional detection methods—such as blacklist filtering, signature-based analysis, and keyword heuristics—proved largely ineffective against these dynamic threats [8]. Many contemporary models relied on narrow feature sets, focusing solely on email content while ignoring stylometric and metadata signals [9]. Proprietary datasets further limited generalizability, and static detection systems struggled to adapt to the evolving tactics of attackers [10]. As phishing strategies continued to evolve, detection systems were perpetually playing catch-up, creating a persistent

gap in defense capabilities. This highlighted the urgent need for a more comprehensive, adaptive, and resilient approach to phishing detection—one that accounted for the origin and complexity of AI-generated threats.

To address this gap, the present study developed a multi-modal origin classification framework that integrated three categories of distinguishing features specifically text-based features, content-based features and metadata attributes [11] [12][13][[14][7]. These features were extracted and analyzed to identify patterns that differentiated AI-generated from human-generated phishing emails. Machine learning models—including SVM, XGBoost, and Logistic Regression—were designed and evaluated using precision, recall, F1-score, and AUC to determine classification performance. Moreover, to ensure resilience against evolving threats, the study assessed adversarial robustness using evasion strategies such as the Fast Gradient Sign Method (FGSM). Through this comprehensive approach, the research advanced phishing detection by addressing origin classification and enhancing model reliability under adversarial conditions.

**2) METHODS AND METHODOLOGY:**

This study employed a quantitative, multi-modal methodology to classify phishing emails by origin, specifically distinguishing between human-generated and AI-generated instances. The analytical framework was structured around three core phases—exploratory, predictive, and diagnostic analysis—each aligned with a distinct research objective. Statistical tests were conducted using Microsoft Excel's Analysis ToolPak, while machine learning workflows were executed in Google Colab.

The dataset comprised balanced samples of the Jose Nazario phishing email dataset and synthetic AI generated phishing emails using GPT 4 OpenAI. Features were extracted across three modalities. Text-based (e.g., stylometric and lexical indicators), content-based (e.g., hyperlink presence, urgency phrases), and metadata-based (e.g., SPF status, timestamp entropy). Preprocessing steps included standardization, encoding, and imputation, with continuous variables assessed for normality to guide statistical test selection.

Exploratory analysis was conducted to identify statistically significant differences between email origins across modalities. Independent sample t-tests were applied to numerical features in the text and content modalities to assess mean differences between groups, as illustrated in equation 1.

$$T = \frac{\overline{X}_1 + \overline{X}_2}{\sqrt{\left[\frac{(n_1 - 1)\, S_1^2 + (n_1 + 1) S_2^2}{n_1 + n_2 - 2}\right]\left[\frac{n_1 + n_2}{n_1 \times n_2}\right]}}$$

Equation 1. Independent Sample T-Test

For binary content-based features, the chi-squared test was used to evaluate independence between feature presence and email origin as seen in equation 2,

$$x^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Equation 2. Chi-Squared Test

Metadata features, which were either non-normally distributed or ordinal, were analyzed using the Mann–Whitney U test see equation 3,

$$U = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - \sum_{i=1}^{n_1} R_i$$

Equation 3. Mann Whitney U Test

These statistical tests provided foundational insights into feature distributions and were executed using Excel's Analysis ToolPak to ensure consistency across modalities.

To identify the most informative features for origin classification, Mutual Information (MI) was employed as a non-parametric feature selection technique. MI quantifies the shared information between a feature and the target class, capturing both linear and non-linear dependencies without assuming normality or linearity. MI scores were computed using the mutual_info_classif function from the

sklearn.feature_selection module, with features discretized where necessary to ensure compatibility. The mathematical formulation is presented in equation 4. Top-ranked features were retained for model training and statistical comparison

$$I(X;Y) = \sum_{x \epsilon X} \sum_{y \mathcal{E} Y} p(x,y) \cdot \log \left( \frac{p(x,y)}{p(x) \cdot p(y)} \right)$$

Equation 3. Mutual Information

Table 1. Model Performance Metrics

| **Metric** | **Equation** |
|---|---|
| Accuracy | $\frac{TP + TN}{TP + TN + FP + FN}$ |
| Precision | $\frac{TP}{TP + FP}$ |
| Recall | $\frac{TP}{TP + FN}$ |
| F1 Score | $2\frac{Precision\ Recall}{Precision + Recall}$ |

Table 1 summarizes the mathematical definitions of the metrics used in the study. Predictive analysis was conducted to design and evaluate machine learning models for classifying phishing emails by origin. Seven algorithms were trained: Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Random Forest, XGBoost, LightGBM, Extra Trees, and Logistic Regression. All models were trained under a consistent preprocessing and stratified k-fold cross-validation framework. Performance was evaluated using accuracy, precision, recall, and F1-score, with metrics averaged across folds to reduce bias from any single data split. Ablation studies were also conducted to compare single-modality, pairwise, and fused feature configurations.

To assess model robustness under adversarial conditions, diagnostic analysis was performed using the Fast Gradient Sign Method (FGSM). Controlled perturbations were applied to the test set features at varying $\varepsilon$ values ranging from 0.001 to 0.5. The impact of these perturbations on classification accuracy was recorded, and the attack success rate was calculated as the proportion of originally correct classifications that became incorrect post-perturbation. This analysis quantified model vulnerability and informed recommendations for more resilient phishing detection systems.

**3] RESULTS:**

Identifying distinguishing features between AI-generated and human-generated phishing email statistical analyses were conducted across text-based, content-based, and metadata modalities.

Table 2. Key Distinguishing Features of AI generated and Human Generated Phishing Email

| **Modality** | **Feature Type** | **Top Features** | **AI vs Human Mean (SD)** | **Test Statistic** | **p-value** |
|---|---|---|---|---|---|
| Text | Stylometric | Lexical Diversity | 0.79 (0.06) vs 0.83 (0.21) | t = -9.00 | < .001 |
| | Psycholinguistic | Personalization Markers | 1.79 (0.60) vs 0.82 (0.84) | t = 49.36 | < .001 |
| | Semantic | NER_PERSON Count | 2.31 (1.24) vs 0.76 (2.80) | t = 26.67 | < .001 |
| Content | Structural | HTML Size (bytes) | 644.43 vs 5018.71 | t = -32.69 | < .001 |
| | Semantic | URL Entropy | 4.24 vs 2.57 | t = 40.07 | < .001 |
| Metadata | Header-level | Auth Fail Score | | U = 7,597,740 | < .001 |

| | | | | | |
|---|---|---|---|---|---|
| | Routing | Message ID Dot Count | | U = 1,868,160 | < .001 |

Table 2 provided the most discriminative features across modalities. It is revealed that there is a distinct behavioral signatures between AI-generated and human-generated phishing emails. In the text modality, lexical diversity emerged as a key differentiator, with AI emails exhibiting lower variability (M = 0.79, SD = 0.06) compared to human phishing (M = 0.83, SD = 0.21), suggesting template-driven phrasing. Named entity recognition (NER_PERSON) counts were significantly higher in AI emails (M = 2.31, SD = 1.24), indicating overuse of personal names to simulate personalization. Psycholinguistic markers such as politeness and imperative verbs were also more frequent in AI samples, reflecting customer-facing tone and urgency scripting. The overall features extracted from the text based can be found in appendix A. In the content modality, URL entropy was notably higher in AI-generated emails (M = 4.24) than in human-crafted ones (M = 2.57), consistent with templated link structures. HTML size and structure further distinguished origins, with human phishing using larger, more complex layouts to mimic legitimate interfaces. To have a more detailed about the extracted features for the content based, please see appendix B. Metadata features such as authentication failure scores and message ID dot counts were significantly elevated in human phishing, revealing spoofed routing paths and header manipulation. A more detailed description for the metadata features can be seen on appendix C. These findings not only validate the feature engineering process but also provide interpretable cues for origin-aware classification and downstream adversarial assessment.

Table 3. Performance Metrics of Models

| **Feature Composition** | **# Features** | **Classifiers** | **Accuracy** | **F1-Score** | **ROC–AUC** |
|---|---|---|---|---|---|
| Text only | 51 | Extra Trees | 0.999 | 0.999 | 0.999 |
| Content only | 20 | Random Forest | 1 | 1 | 1 |
| Metadata only | 10 | Logistic Regression | 1 | 1 | 1 |
| Text + Content | 71 | XGBoost | 1 | 1 | 1 |
| Text + Metadata | 61 | LightGBM | 1 | 1 | 1 |
| Content + Metadata | 30 | Random Forest | 1 | 1 | 1 |
| Text + Content + Metadata | 114 | LR (Adversarial), LightGBM, XGBoost, MLP, RF, ET | 1 | 1 | 1 |

Table 3 presents a consolidated overview of classifier performance across single, dual, and tri-modal feature configurations. In unimodal setups, metadata features achieved perfect classification using only 10 indicators, with Logistic Regression yielding an accuracy, F1-score, and ROC–AUC of 1.000 across all folds. This highlights the discriminative strength of header-level attributes and supports metadata's role in parsimonious modeling. Content-based models reached flawless performance with 20 features, with Random Forest leading, while text-based models required 51 features to approach similar accuracy, with Extra Trees achieving 0.999 across all metrics. A more detailed result of each single modality result can be found in appendix D. Dual-modal fusion further enhanced results: all combinations—Text + Content (71 features), Text + Metadata (61), and Content + Metadata (30)—achieved perfect scores (1.000) across classifiers, with XGBoost and LightGBM consistently outperforming baselines. These results suggest that cross-modal interactions between stylometric, semantic, and structural cues contribute meaningfully to phishing origin classification. Appendix E provide a detailed result of each modality combination. In the multi-modal configuration (114 features), all classifiers—including LR, RF, ET, XGBoost, LightGBM, and MLP—achieved ceiling-level metrics, confirming the robustness and generalizability of the integrated

feature set. Logistic Regression was retained for adversarial assessment due to its interpretability, baseline stability, and consistent performance across modalities. Appendix F shows how these models performed. These findings validate the effectiveness of engineered features and support the use of multi-modal fusion for origin-aware phishing detection, with parsimony preserved even in high-dimensional setups.

To evaluate the resilience of phishing origin classification under adversarial conditions, Logistic Regression (LR) was subjected to gradient-based evasion attacks using the Fast Gradient Sign Method (FGSM). The model had been trained exclusively on the top 10 metadata features identified in table 2, which previously yielded perfect classification performance under clean conditions. Prior to adversarial perturbation, LR achieved flawless classification on the test set, with an accuracy, F1-score, and ROC–AUC of 1.000. The confusion matrix confirmed zero misclassifications across 1,668 samples, validating the discriminative strength of metadata features and the model's baseline reliability.

Table 4. Adversarial Robustness Evaluation of LR

| Epsilon | LR Accuracy | LR Misclassification Rate | Misclassified (LR) |
|---|---|---|---|
| 0.001 | 0.7998 | 0.2002 | 334 / 1668 |
| 0.005 | 0.7974 | 0.2026 | 338 / 1668 |
| 0.01 | 0.7554 | 0.2446 | 408 / 1668 |
| 0.05 | 0.7086 | 0.2914 | 486 / 1668 |
| 0.1 | 0.6517 | 0.3483 | 581 / 1668 |
| 0.2 | 0.5695 | 0.4305 | 718 / 1668 |
| 0.3 | 0.4994 | 0.5006 | 835 / 1668 |
| 0.5 | 0.06 | 0.94 | 1568 / 1668 |

Table 4 illustrates the Logistic regression nnder FGSM perturbations, model performance declined sharply as epsilon (ϵ) increased. At minimal perturbation (ϵ = 0.001), accuracy dropped to 79.98%, with 334 misclassified samples. At moderate levels (ϵ = 0.05), accuracy fell to 70.86%, and at ϵ = 0.3, the model misclassified over half the samples (Accuracy = 49.94%). At maximum perturbation (ϵ = 0.5), accuracy collapsed to 6.00%, with 1,568 of 1,668 samples misclassified

## 4] CONCLUSION:

The findings demonstrate that multi-modal feature sets can effectively differentiate AI-generated from human-crafted phishing emails. Distinct patterns in stylometric, semantic, structural, and metadata indicators offer a quantifiable basis for origin classification, underscoring the importance of incorporating diverse modalities into detection systems. These distinctions reflect the evolving nature of phishing tactics and highlight the need for adaptive, origin-aware cybersecurity solutions.

The evaluation of machine learning models revealed that both individual and fused feature sets consistently achieved near-perfect classification performance. Ensemble algorithms, in particular, demonstrated high accuracy and efficiency across modalities, validating the advantage of integrating heterogeneous data cues. These results suggest that enterprise-level email filtering systems should adopt multi-modal architectures to enhance threat detection capabilities and reduce false negatives in operational environments.

Diagnostic analysis exposed a critical vulnerability in model robustness under adversarial conditions. Even minor perturbations introduced via the Fast Gradient Sign Method led to significant performance degradation, particularly in linear classifiers. This finding emphasizes that high accuracy alone is insufficient for real-world deployment and calls for the integration of adversarial resilience mechanisms and continuous robustness evaluation in cybersecurity pipelines.

## 5] Acknowledgement:

We thank Arnemie Gayyed and Natividad Concepcion for their contribution to this work. Special thanks to the University of the Cordilleras for their institutional support, research resources, and academic guidance throughout this study.

**6] Funding Statement:** There is no fund received for this article.

**7] Data Availability:** The data that support the findings of this study are available from https://monkey.org/%7Ejose/phishing/ for the human-generated phishing emails. While the AI-generated phishing emails are available upon request.

**8] Conflict of interest:** The authors declare that there is no conflict of interest**.**

**9] REFERENCES:**

[1] Banerjee, A., Agrawal, C., & Chaturvedi, S. (2024). Comprehensive methodology for phishing detection and awareness: Unravelling the intricacies of cyber threats. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4851643
[2] Gopalsamy, Mani. (2024). Identification And Classification Of Phishing Emails Based on Machine Learning Techniques To Improvise Cyber security. 10. 47-55.
[3] Kumar, S, Menezes, A., Giri, S., & Kotikela, S. (2024). What the phish! Effects of AI on phishing attacks and defense. 4(1), 218–226. https://doi.org/10.34190/icair.4.1.3224
[4] Heiding, F., Hua, J., Wang, P., & Lutchkus, P. (2024). How effective are large language models in detecting phishing emails? Issues in Information Systems, 25(3), 327–341.
[5] Eze, C. S., & Shamir, L. (2024). Analysis and prevention of AI-based phishng email attacks. Electronics, 13(10), 1839. https://doi.org/10.3390/electronics13101839
[6] ANDRIU, A.-V. (2023). Adaptive phishing detection: Harnessing the power of artificial intelligence for enhanced email security. Romanian Cyber Security Journal, 5(1), 3–9. https://doi.org/10.54851/v5i1y202301
[7] Zheng, S., & Becker, I. (2022). Presenting suspicious details in user-facing e-mail headers does not improve phishing detection. USENIX SOUPS 2022. https://www.usenix.org/system/files/soups2022-zheng_1.pdf
[8] Thakur, K., Ali, M. L., Obaidat, M. A., & Kamruzzaman, A. (2023). A systematic review on deep-learning-based phishing email detection. Electronics, 12(21), 4545. https://doi.org/10.3390/electronics12210045
[9] Doshi, J., Parmar, K., Sanghavi, R., & Shekokar, N. (2023). A comprehensive dual-layer architecture for phishing and spam email detection. Computers & Security, 133, 103378. https://doi.org/10.1016/j.cose.2023.103378
[10] Al-Yozbaky, R. S., & Alanezi, M. (2023). A review of different content-based phishing email detection methods. In 2023 9th International Engineering Conference on Sustainable Technology and Development (IEC) (pp. 20–25). IEEE. https://doi.org/10.1109/iec57380.2023.10438812
[11] Opara, C., Modesti, P., & Golightly, L. (2025). Evaluating spam filters and stylometric detection of AI-generated phishing emails. Expert Systems With Applications, 276, 127044. https://doi.org/10.1016/j.eswa.2025.127044
[12] Qi, Q., Luo, Y., Xu, Y., Guo, W., & Fang, Y. (2025). SpearBot: Leveraging large language models in a generative-critique framework for spear-phishing email generation. Information Fusion, 103176. https://doi.org/10.1016/j.inffus.2025.103176
[13] Murti, Y., & Naveen, P. (2023). Machine learning algorithms for phishing email detection. Journal of Logistics Informatics and Service Science, 10(2), 17. https://doi.org/10.33168/jliss.2023.0217
[14] Kulkarni, M., Kumar, S., Panjwani, Y., Mohana, N., Moharir, M., Kumar, A. R. A., & Baskaran, E. (2024). Mitigating email phishing: Analytical framework, simulation models, and preventive measures. In 2024 International Conference on Computer Science, Engineering and Applications (pp. 1459–1464). IEEE. https://doi.org/10.1109/iccsp60870.2024.10543325

**10] Miscellaneous:**

Appendix A. Text Modality Features

| No | Feature | AI | | HUMAN | | | |
|---|---|---|---|---|---|---|---|
| | | Mean | SD | Mean | SD | Test Statistic | P-value |
| 1 | Aggressiveness_Marker_Count | 0.56 | 0.5 | 0.24 | 0.49 | 24.35 | 0 |
| 3 | Email Body_past_tense_frequency | 0.02 | 0.02 | 0.01 | 0.02 | 19.19 | 0 |
| 4 | Email Body_avg_sentence_length | 16.32 | 4.8 | 12.39 | 10.88 | 17.39 | 0 |
| 5 | Lexical_Diversity | 0.79 | 0.06 | 0.83 | 0.21 | -9 | 0 |
| 6 | Dale_Chall | 12.2 | 1 | 12.18 | 3.81 | 0.19 | 0.8519 |
| 7 | Unique_Word_Count | 29.13 | 4.01 | 28.68 | 35.52 | 0.66 | 0.512 |
| 8 | exclamation_count | 0 | 0 | 0.29 | 5.14 | -2.98 | 0.0029 |
| 9 | word_length_variation | 2.46 | 0.18 | 2.85 | 3.43 | -6.07 | 0 |
| 10 | Urgency_Markers_Count | 0.33 | 0.47 | 0.11 | 0.34 | 19.52 | 0 |
| 11 | Lemma_Diversity | 0.8 | 0.07 | 0.82 | 0.22 | -4.34 | 0 |
| 12 | Email Body_complex_word_count | 9.11 | 3.15 | 10.25 | 16.32 | -3.6 | 0.0003 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 13 | comma_count | 4.15 | 1.17 | 3.42 | 27.93 | 1.36 | 0.1737 |
| 14 | SMOG_Index | 10.84 | 1.11 | 11.19 | 3.15 | -5.5 | 0 |
| 15 | Technical_Jargon_Count | 1.43 | 0.86 | 0.6 | 0.81 | 37.18 | 0 |
| 16 | Conditional_Phrases_Count | 0.37 | 0.48 | 0.2 | 0.42 | 14.03 | 0 |
| 17 | Email Body_cardinal_number_frequency | 0 | 0 | 0 | 0 | -7.65 | 0 |
| 18 | word_count | 70.46 | 11.67 | 80.2 | 167.2 | -3.06 | 0.0022 |
| 19 | Personalization_Markers_Count | 1.79 | 0.6 | 0.82 | 0.84 | 49.36 | 0 |
| 20 | Politeness_Markers_Count | 1.1 | 0.83 | 0.63 | 0.82 | 21.33 | 0 |
| 21 | Flesch_Reading_Ease | 36.46 | 10.5 | 34.21 | 55.93 | 2.08 | 0.0376 |
| 22 | Email Body_bigram_count | 49.93 | 6.41 | 57.22 | 73.66 | -5.2 | 0 |
| 23 | quotation_count | 0 | 0 | 0.77 | 13.3 | -3.07 | 0.0022 |
| 24 | Email Body_first_person_pronoun_count | 1.25 | 0.7 | 1.32 | 2.29 | -1.53 | 0.1265 |
| 25 | total_punct_counts | 31.8 | 7.26 | 22.06 | 109 | 4.7 | 0 |
| 26 | Email Body_pronoun_density | 0.08 | 0.02 | 0.06 | 0.06 | 18.31 | 0 |
| 27 | punctuation_variety | 4.87 | 0.38 | 3.49 | 1.96 | 36.58 | 0 |
| 28 | Imperative_Verbs_Count | 1.33 | 1.02 | 0.84 | 0.98 | 18.48 | 0 |
| 29 | avg_syllables_per_word | 1.6 | 0.08 | 1.61 | 0.39 | -1.82 | 0.069 |
| 30 | Email Body_second_person_pronoun_count | 3.22 | 1.3 | 3.74 | 4.87 | -5.41 | 0 |
| 31 | Email Body_sentence_count | 3.64 | 1.23 | 4.44 | 4.7 | -8.66 | 0 |
| 32 | dash_count | 3.83 | 2.95 | 5.38 | 45.81 | -1.78 | 0.0754 |
| 33 | Email Body_pronoun_count | 4.48 | 1.4 | 5.54 | 7.29 | -7.55 | 0 |
| 34 | Promotional_Word_Count | 0 | 0.05 | 0.08 | 0.27 | -13.95 | 0 |
| 35 | avg_word_length | 5.28 | 0.27 | 5.11 | 1.67 | 5.33 | 0 |
| 36 | character_count | 472.4 | 74.15 | 574.3 | 1253 | -4.28 | 0 |
| 37 | Email Body_function_word_density | 0.23 | 0.03 | 0.27 | 0.12 | -16.64 | 0 |
| 38 | colon_count | 7.7 | 1.96 | 2.32 | 16 | 17.59 | 0 |
| 39 | semicolon_count | 0 | 0 | 1.2 | 18.1 | -3.49 | 0.0005 |
| 40 | Gunning_Fog_Index | 11.38 | 2.02 | 12.42 | 4.87 | -10.42 | 0 |
| 41 | Modal_Verb_Count | 0.27 | 0.45 | 0.94 | 1.38 | -24.2 | 0 |
| 42 | Email Body_trigram_count | 51.03 | 6.72 | 59.18 | 79.4 | -5.39 | 0 |
| 43 | Email Body_pos_diversity | 0.27 | 0.04 | 0.37 | 0.24 | -22.65 | 0 |
| 44 | Coleman_Liau | 17.53 | 2.1 | 16.58 | 25.54 | 1.95 | 0.051 |
| 45 | Email Body_avg_sentiment | 0.34 | 0.31 | 0.08 | 0.16 | 40.33 | 0 |
| 46 | Uncertainty_Adverbs_Count | 0 | 0 | 0.01 | 0.07 | -3.88 | 0.0001 |
| 47 | Email | 12.8 | 2.62 | 20.7 | 26.9 | -15.5 | 0 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Body_functional_word_count | | | 6 | 7 | | |
| 48 | NER_PERSON_Count | 2.31 | 1.24 | 0.76 | 2.8 | 26.67 | 0 |
| 49 | NER_ORG_Count | 2.36 | 1.27 | 2.02 | 6.36 | 2.73 | 0.0064 |
| 50 | NER_GPE_Count | 0.16 | 0.43 | 0.33 | 1.14 | -7.66 | 0 |
| 51 | CTA_Phrase_Count | 0 | 0 | 0.17 | 0.41 | -22.36 | 0 |
| 52 | Hedging_Word_Count | 0 | 0 | 0.06 | 0.31 | -10.29 | 0 |
| 53 | Stopword_Ratio | 0.22 | 0.04 | 0.25 | 0.12 | -10.97 | 0 |

Appendix B. Content Modality Features

| No | Feature | AI | | Human | | Test Statistic | P-value |
|---|---|---|---|---|---|---|---|
| | | Mean | SD | Mean | SD | | |
| 1 | avg_url_length | 33.65 | 4 | 52.1 | 101 | -9.63 | 0 |
| 3 | num_iframes | 0 | 0 | 0 | 0.02 | -1 | 0.3174 |
| 4 | avg_attachment_filename_entropy | 0.43 | 1.08 | 0.4 | 1.2 | 0.82 | 0.4094 |
| 5 | avg_url_entropy | 4.24 | 0.15 | 2.57 | 2.2 | 40.07 | 0 |
| 6 | num_images_in_html | 0.63 | 0.7 | 1.62 | 3.91 | -13.07 | 0 |
| 7 | html_size_bytes | 644.4 | 186.6 | 5019 | 7053 | -32.69 | 0 |
| 8 | num_urls | 2.08 | 0.96 | 1.62 | 2.62 | 8.72 | 0 |
| 9 | num_css_files | 0 | 0 | 0.01 | 0.15 | -4.68 | 0 |
| 10 | density_of_links_in_html | 0.02 | 0.02 | 0.02 | 0.03 | -6.19 | 0 |
| 11 | num_forms | 0 | 0 | 0 | 0.04 | -2.24 | 0.0253 |
| 12 | num_input_fields | 0 | 0 | 0 | 0.05 | -2.45 | 0.0143 |
| 13 | num_attachments | 0.14 | 0.34 | 0.1 | 0.3 | 3.72 | 0.0002 |
| 14 | avg_num_hyphens_in_url | 0.32 | 0.38 | 0.71 | 2.56 | -7.96 | 0 |
| 15 | num_scripts | 0 | 0 | 0.01 | 0.23 | -3.27 | 0.0011 |
| 16 | max_url_length | 36.41 | 5.78 | 71 | 159.1 | -11.45 | 0 |
| 17 | avg_num_dots_in_url | 1.4 | 0.31 | 1.5 | 2.05 | -2.54 | 0.011 |
| 18 | avg_num_slashes_in_url | 3.27 | 0.28 | 2.56 | 3.88 | 9.49 | 0 |
| 19 | avg_num_digits_in_url | 0.97 | 0.89 | 5.97 | 20.3 | -12.99 | 0 |
| 20 | avg_attachment_filename_length | 1.5 | 3.83 | 2.89 | 9.57 | -7.12 | 0 |
| 21 | Link_Mismatch_Count | 0 | 0 | 0.01 | 0.31 | -1.48 | 0.1397 |

| No | Featur | AI Presence % | Human Presenc | Statis | Test S | P-valu | Interp |
|---|---|---|---|---|---|---|---|
| 1 | has an | 0.00% | 5.00% | Chi-Sq | 140.5 | 0 | S |
| 2 | has an | 0.18% | 13.71% | Chi-Sq | 391.5 | 0 | S |
| 3 | has at | 13.60% | 10.36% | Chi-Sq | 13.51 | 2E-04 | S |
| 4 | input | 0.00% | 0.00% | Fisher | - | 1 | NS |
| 5 | has ur | 100.00% | 58.92% | Chi-Sq | 1435 | 0 | S |
| 6 | has an | 0.00% | 20.47% | Chi-Sq | 631.6 | 0 | S |
| 7 | has ba | 0.00% | 0.32% | Chi-Sq | 7.12 | 0.008 | S |
| 8 | has an | 0.00% | 0.36% | Chi-Sq | 8.11 | 0.004 | S |
| 9 | has an | 0.00% | 0.00% | Fisher | - | 1 | NS |
| 10 | has an | 4.39% | 8.02% | Chi-Sq | 30.9 | 0 | S |
| 11 | has ht | 50.47% | 93.60% | Chi-Sq | 1281 | 0 | S |
| 12 | has ex | 0.00% | 0.14% | Chi-Sq | 2.25 | 0.134 | NS |
| 13 | has ur | 0.00% | 0.00% | Fisher | - | 1 | NS |
| 14 | has an | 9.24% | 2.77% | Chi-Sq | 102.1 | 0 | S |
| 15 | has ex | 31.73% | 29.24% | Chi-Sq | 3.92 | 0.048 | S |
| 16 | has hi | 0.00% | 4.53% | Chi-Sq | 126.9 | 0 | S |
| 17 | has an | 0.00% | 7.45% | Chi-Sq | 212.9 | 0 | S |
| 18 | has fo | 0.00% | 0.18% | Chi-Sq | 3.2 | 0.074 | NS |

Appendix C. Metadata Modality Features

| No | Feature | Test Statistic | P-value | Interpretation |
|---|---|---|---|---|
| 1 | auth_fail_score | 7597740 | 0 | S |
| 2 | bcc | 4937280 | 1.72E-195 | S |
| 3 | cc | 5600310 | 0 | S |
| 4 | date | 3862810 | 0.980192943 | NS |
| 5 | email_hour | 3745006 | 0.046216992 | S |
| 6 | from_subdomain_count | 159850 | 0 | S |
| 7 | multiple_attachments | 3864200 | 1 | NS |
| 8 | return_path_dot_count | 1087194 | 0 | S |
| 9 | reply_to_dot_count | 5382500 | 1.44E-198 | S |
| 10 | user_agent_dot_count | 7678900.5 | 0 | S |
| 11 | x_mailer_dot_count | 6655186.5 | 0 | S |
| 12 | attachment_type_length | 2458910 | 3.98E-223 | S |
| 13 | attachment_type_dot_count | 3864200 | 1 | NS |
| 14 | message_id_length | 4702036.5 | 1.03E-44 | S |
| 15 | message_id_dot_count | 1868160 | 3.43E-295 | S |
| 16 | return_path_length | 1162186 | 0 | S |
| 17 | user_agent_length | 7727254 | 0 | S |
| 18 | x_mailer_length | 7471678 | 0 | S |
| 19 | ip_address_last_octet | 3857250 | 0.893566336 | NS |
| 20 | auth_fail_score_label | These are the features that cannot be computed through statistics but utilized in training the model | | |
| 21 | country | | | |
| 22 | dkim_fail | | | |
| 23 | dkim_result | | | |
| 24 | dmarc_fail | | | |
| 25 | dmarc_result | | | |
| 26 | email_client | | | |
| 27 | email_dayofweek | | | |
| 28 | email_weekend | | | |
| 29 | from_domain | | | |
| 30 | from_is_free | | | |
| 31 | from_tld | | | |
| 32 | is_dkim_aligned | | | |
| 33 | is_spf_aligned | | | |
| 34 | is_urgent | | | |
| 35 | priority | | | |
| 36 | reply_mismatch | | | |
| 37 | reply_to_domain | | | |
| 38 | spf_fail | | | |
| 39 | spf_result | | | |
| 40 | timezone | | | |
| 41 | user_agent_missing | | | |
| 42 | user_agent_scripted | | | |
| 43 | has_attachment_flag | | | |
| 44 | to_domain | | | |

Appendix D. Performance of Models Using Single Modality

**Text Modality**

| Model | cv_score | K | f1 | ROC_AUC | Accuracy | Precision | Recall |
|---|---|---|---|---|---|---|---|
| SVM | 0.9964 | 51 | 0.9982 | 0.9991 | 0.9982 | 0.9982 | 0.9982 |
| LR | 0.9936 | 51 | 0.9922 | 0.9985 | 0.9922 | 0.9923 | 0.9922 |
| RF | 0.9987 | 51 | 0.9982 | 1 | 0.9982 | 0.9982 | 0.9982 |
| ET | 0.9997 | 51 | 0.9994 | 1 | 0.9994 | 0.9994 | 0.9994 |
| LGBM | 0.9982 | 51 | 0.9994 | 0.9999 | 0.9994 | 0.9994 | 0.9994 |
| XGB | 0.9974 | 51 | 0.9988 | 1 | 0.9988 | 0.9988 | 0.9988 |
| MLP | 0.9972 | 20 | 0.9976 | 0.9988 | 0.9976 | 0.9976 | 0.9976 |

**Content Modality**

| Model | cv_score | K | f1 | ROC_AUC | Accuracy | Precision | Recall |
|---|---|---|---|---|---|---|---|
| SVM | 0.9943 | 20 | 0.9946 | 0.9999 | 0.9946 | 0.9947 | 0.9946 |
| LR | 0.9895 | 38 | 0.988 | 0.9996 | 0.988 | 0.9883 | 0.988 |
| RF | 0.999 | 20 | 1 | 1 | 1 | 1 | 1 |
| ET | 0.999 | 20 | 0.9994 | 1 | 0.9994 | 0.9994 | 0.9994 |
| LGBM | 0.9982 | 20 | 1 | 1 | 1 | 1 | 1 |
| XGB | 0.9982 | 20 | 1 | 1 | 1 | 1 | 1 |
| MLP | 0.9972 | 20 | 0.9976 | 1 | 0.9976 | 0.9976 | 0.9976 |

**Metadata Modality**

| Model | cv_score | K | f1 | ROC_AUC | Accuracy | Precision | Recall |
|---|---|---|---|---|---|---|---|
| SVM | 1 | 10 | 1 | 1 | 1 | 1 | 1 |
| LR | 1 | 10 | 1 | 1 | 1 | 1 | 1 |
| RF | 1 | 10 | 1 | 1 | 1 | 1 | 1 |
| ET | 1 | 10 | 1 | 1 | 1 | 1 | 1 |
| LGBM | 0.9992 | 10 | 1 | 1 | 1 | 1 | 1 |
| XGB | 0.9995 | 10 | 1 | 1 | 1 | 1 | 1 |
| MLP | 1 | 10 | 1 | 1 | 1 | 1 | 1 |

Appendix E. Performance of Models Using Dual Modality

Text + Content

| Model | cv_score | K | f1 | ROC_AUC | Accuracy | Precision | Recall |
|---|---|---|---|---|---|---|---|
| SVM | 0.9979 | 89 | 0.9964 | 0.9987 | 0.9964 | 0.9964 | 0.9964 |
| LR | 0.9979 | 89 | 0.9964 | 0.9987 | 0.9964 | 0.9964 | 0.9964 |
| RF | 0.9997 | 89 | 1 | 1 | 1 | 1 | 1 |
| ET | 1 | 89 | 1 | 1 | 1 | 1 | 1 |
| LGBM | 0.9987 | 89 | 0.9976 | 1 | 0.9976 | 0.9976 | 0.9976 |
| XGB | 0.9992 | 89 | 0.9982 | 0.9999 | 0.9982 | 0.9982 | 0.9982 |
| MLP | 0.999 | 89 | 0.9976 | 0.9996 | 0.9976 | 0.9976 | 0.9976 |

**Text + Metadata**

| Model | cv_score | K | f1 | ROC_AUC | Accuracy | Precision | Recall |
|---|---|---|---|---|---|---|---|
| SVM | 0.999 | 76 | 0.9982 | 1 | 0.9982 | 0.9982 | 0.9982 |
| LR | 0.9987 | 76 | 0.9976 | 1 | 0.9976 | 0.9976 | 0.9976 |
| RF | 0.9995 | 20 | 0.9994 | 1 | 0.9994 | 0.9994 | 0.9994 |
| ET | 0.9992 | 10 | 0.9988 | 1 | 0.9988 | 0.9988 | 0.9988 |
| LGBM | 0.999 | 10 | 0.9994 | 1 | 0.9994 | 0.9994 | 0.9994 |
| XGB | 0.999 | 10 | 0.9994 | 1 | 0.9994 | 0.9994 | 0.9994 |
| MLP | 0.999 | 76 | 0.9976 | 1 | 0.9976 | 0.9976 | 0.9976 |

**Content + Metadata**

| Model | cv_score | K | f1 | ROC_AUC | Accuracy | Precision | Recall |
|---|---|---|---|---|---|---|---|

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| SVM | 0.9997 | 63 | 1 | 1 | 1 | 1 | 1 |
| LR | 0.9997 | 63 | 1 | 1 | 1 | 1 | 1 |
| RF | 1 | 63 | 1 | 1 | 1 | 1 | 1 |
| ET | 0.9995 | 63 | 1 | 1 | 1 | 1 | 1 |
| LGBM | 0.999 | 63 | 0.9988 | 1 | 0.9988 | 0.9988 | 0.9988 |
| XGB | 0.9992 | 63 | 1 | 1 | 1 | 1 | 1 |
| MLP | 0.9997 | 63 | 1 | 1 | 1 | 1 | 1 |

**Text+Content+Metadata**

| **Model** | **cv_score** | **K** | **f1** | **ROC_AUC** | **Accuracy** | **Precision** | **Recall** |
|---|---|---|---|---|---|---|---|
| SVM | 0.9997 | 114 | 0.9994 | 1 | 0.9994 | 0.9994 | 0.9994 |
| LR | 0.9995 | 114 | 1 | 1 | 1 | 1 | 1 |
| RF | 1 | 114 | 1 | 1 | 1 | 1 | 1 |
| ET | 0.9997 | 114 | 1 | 1 | 1 | 1 | 1 |
| LGBM | 0.9992 | 114 | 1 | 1 | 1 | 1 | 1 |
| XGB | 0.999 | 114 | 1 | 1 | 1 | 1 | 1 |
| MLP | 0.9997 | 114 | 0.9994 | 0.9999 | 0.9994 | 0.9994 | 0.9994 |